Nonlinear Dynamics, Psychology, and Life Sciences, Vol. 10, No. 4, pp. 485-488.

Book Review

Advances in Minimum Description Length: Theory and Applications, Edited by Peter D. Grünwald, In Jae Myung and Mark A. Pitt. Cambridge, MA: Bradford/MIT Press. x + 444 pp. ISBN 0-262-07262-9.

The concept of Kolmogorov complexity was independently introduced by Solomonoff, Kolmogorov and Chaitin in the 1960s. In 1978 Rissanen introduced modelling by shortest data description, which builds on the foundation work (accessible in Rissanen, 1987, 2001). There have been further developments, and this book reviews and applies them. It is pleasing that one of the areas of application is explicitly psychology, and this book's last three chapters investigate clustering, perception and cognition. The authors have divided the work between themselves, so that the book reads like a coherent set of review papers, grouped into introductory chapters including a valuable elementary tutorial, theoretical advances, applications to learning, and practical applications. It was developed from a workshop on Neural Information Processing Systems.

To illustrate I repeat the core example that demonstrates how we might learn the underlying laws and regularities in time series that are bits of life histories. If we have three series like

(i) prstprstprstprstprstprst…..n times,
(ii) a completely random binary series of 0s and 1s, and
(iii) 00000010001001100010001010……m terms,

where (iii) has some regularities and some apparent random subseries. The first series can be completely encoded by five symbols, prst *n*, it can be compressed without loss of information into those five symbols, or 2.3219 bits, (ii) cannot be compressed at all, so if it is *m* symbols long its shortest encoding has *m* symbols or $\log_2(m)$ bits, and (iii) lies somewhere between the first two cases. The symbols can be letters, binary numbers or scalar or vector measures in a metric space. The case that needs most theoretical work is obviously (iii). It can be shown that it can be compressed to some length αm, where $0 < \alpha < 1$. The key idea, due first to Solomonoff, is that the ultimate model for a sequence of data may be identified with the shortest computer program that prints the data. There are critical limitations to this idea, it does not follow that a Minimum Data Length (MDL) always exists (Li & Vitányi, 1997), and for short data series the choice of MDL may depend on the symbolic language being used. These difficulties with short series are not due to using numerical or symbolic dynamics, Wolfram (2002, chapter 10) notes that data compression with cellular automata faces similar indeterminacies.

There are links between MDL and the problem of overfitting, where a polynomial of degree (n-1) will always fit without loss a series of n terms, but can be useless for prediction of regularities in an extrapolation of the data. Other important links are to Bayesian statistics, which are extensively discussed, and also to Bayes nets in inference (Pearl, 1988). Essentially MDL is about comparing families of possible models, where we do not know if there exists a 'true' model, and we do not need to know that to proceed. MDL is an active area of research, and extensions and redefinitions of basic Bayesian ideas are being pursued, including links into chaos and multifractal dynamics (Davison & Shiner, 2005; Kennel, Shlens, Abarbanel & Chichilnisky, 2005).

Chapter 5, by Hanson and Fu is an intuitively readable explanation of how to start, with the equation:

Description length Equals Fit plus Geometry.

Fit is an information-theoretic measure corresponding to the number of bits of description length attributable to inaccuracy, and Geometry is the parametric complexity of the model being considered. A full mathematical form of the equation is quite complicated, equations 3.26 and 3.29 are examples (on page 93). A series of worked examples on artificial data ends in section 5.6 with a comparison of Fechner (log) and Stevens (power law) models in psychophysics (European scholars more properly attribute the power law model to Plateau). A fuller treatment of this comparison is given in Pitt et al. (2002). The key finding was that if MDL is used then all data from a model are correctly attributed to that model, but if only goodness-of-fit is used the all data samples were attributed to the power law model, some therefore falsely. Goodness-of-fit is not a sufficient basis for model selection; this point is related to the overfitting question considered previously. However there is a subtle and treacherous circularity buried in this psychophysical example. Neither model is meaningful without specification of its stimulus boundary conditions, and the artificial data have those built in. Psychophysical models do not hold over wide undetermined stimulus ranges, and the appropriate model can be strongly conditional on the ranges used (Eisler, Eisler, & Montgomery, 2004).

Chapter 14 on psychological clustering models turns quickly into consideration of some psychometric similarity models and is attenuated by its uninformed and narrow selection of models. Worse, it treats similarity as a sort of static input-output mapping, when in fact similarity is stimulus range and context dependent (Gregson, 1980) and shows hysteresis; sampling of dynamic psychological processes is still a serious but essentially worthwhile challenge for MDL.

Chapter 15 on perception offers interesting linkages between philosophy of science, the similarity principle in perception (remember Gestalt psychology's 'good form'?), sufficiency in statistics, Bayes and MDL. The idea is developed that perceptual processes rest not on probability of identification, but on coding. Chater's treatment is not mathematical, but goes back to questions that Mach asked, can perception be driven by a search for economy, and economy in MDL is another name for minimal description. Minimal description involves filtering out noise of one sort or another, and one cannot postulate computational process for that which are too complicated for the brain to perform (section 15.4), so we are still faced with the questions of what sorts of MDL are actually used in biological reality.

The last Chapter, 16, is about MDL and cognitive modelling. It focusses heavily on new developments in MDL theory (Section 16.2). This recapitulates and extends the mathematical treatments of some earlier chapters, and treats relations between complexity and nonlinearity in MDL. Again we are faced with performing the trade-off between goodness-fit and model complexity, and the comparisons are made between BIC (Bayesian information criterion) and CV (cross-validation), both of which have their own literature, and their limitations. A worked example comparing the three approaches to three memory retention models is given; it seems best at this state of our knowledge to hang on to all three methods, the only thing that can be safely dumped is significance testing of null hypotheses.

A word of warning to the intending reader; don't try to read the chapters in serial order, unless you are already familiar with stochastic theory, Fisher information, entropy, log likelihood functions, cost-risk analysis, Kullback-Leibler loss, and Bayesian priors, they all come into the picture. The chapters vary in their presuppositions about the reader's erudition even though they complement one another rigorously. Chapter 6 on Algorithmic Statistics and Kolmogorov's Structure Functions gives some valuable insights into the differences between classical statistics and Kolmogorov complexity; the latter ''allows us to precisely formulate and quantify how well a particular model fits a particular piece of data, a matter which formerly was judged impossible''(p. 153). At the same time I am left with the feeling that some of the authors have forgotten the excellent Bayesian principle that one does not start to consider the prior plausibility of a model until one has checked out as much as possible of relevant information about how it describes real-world data.

REFERENCES

Davison, M. & Shiner, J. S. (2005) Extended entropies and disorder. *Advances in Complex Systems, 8,* 125-158.

Eisler, A. D., Eisler, H. & Montgomery, H. (2004). A quantitative model for retrospective subjective duration. *NeuroQuantology, 4,* 263-291.

Gregson, R. A. M. (1980) Model evaluation via stochastic parameter convergence as on-line system identification. *British Journal of Mathematical and Statistical Psychology, 33*, 17-35.

Kennel, M. B., Shlens, J., Abarbanel, H. D. I. & Chichilnisky, E. J. (2005). Estimating entropy rates with Bayesian Confidence Intervals. *Neural Computation, 17*, 1531-1576.

Li, M., & Vitányi, P. (1997) An introduction to Kolmogorov complexity and its applications (2nd edition). New York: Springer Verlag.

Pearl, J. (1988) Probabilistic reasoning in intelligent systems. San Mateo, CA: Morgan Kaufmann.

Pitt, M. A., Myung, I, J. & Zhang, S. (2002) Towards a method of selecting among computational models of cognition. *Psychological Review, 109,* 472-491.

Rissanen, J. (1987) Stochastic complexity. *Journal of the Royal Statistical Society, Series B (Methodological), 49*, 223-239.

Rissanen, J. (2001) Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47,* 1712-1717.

Wolfram, S. (2002). *A new kind of science.* Champaign, IL: Wolfram Media Inc.

— Robert A. M. Gregson
Australian National University
Canberra
e-mail: ramgdd@bigpond.com